BIO392
Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**
Computational Oncogenomics

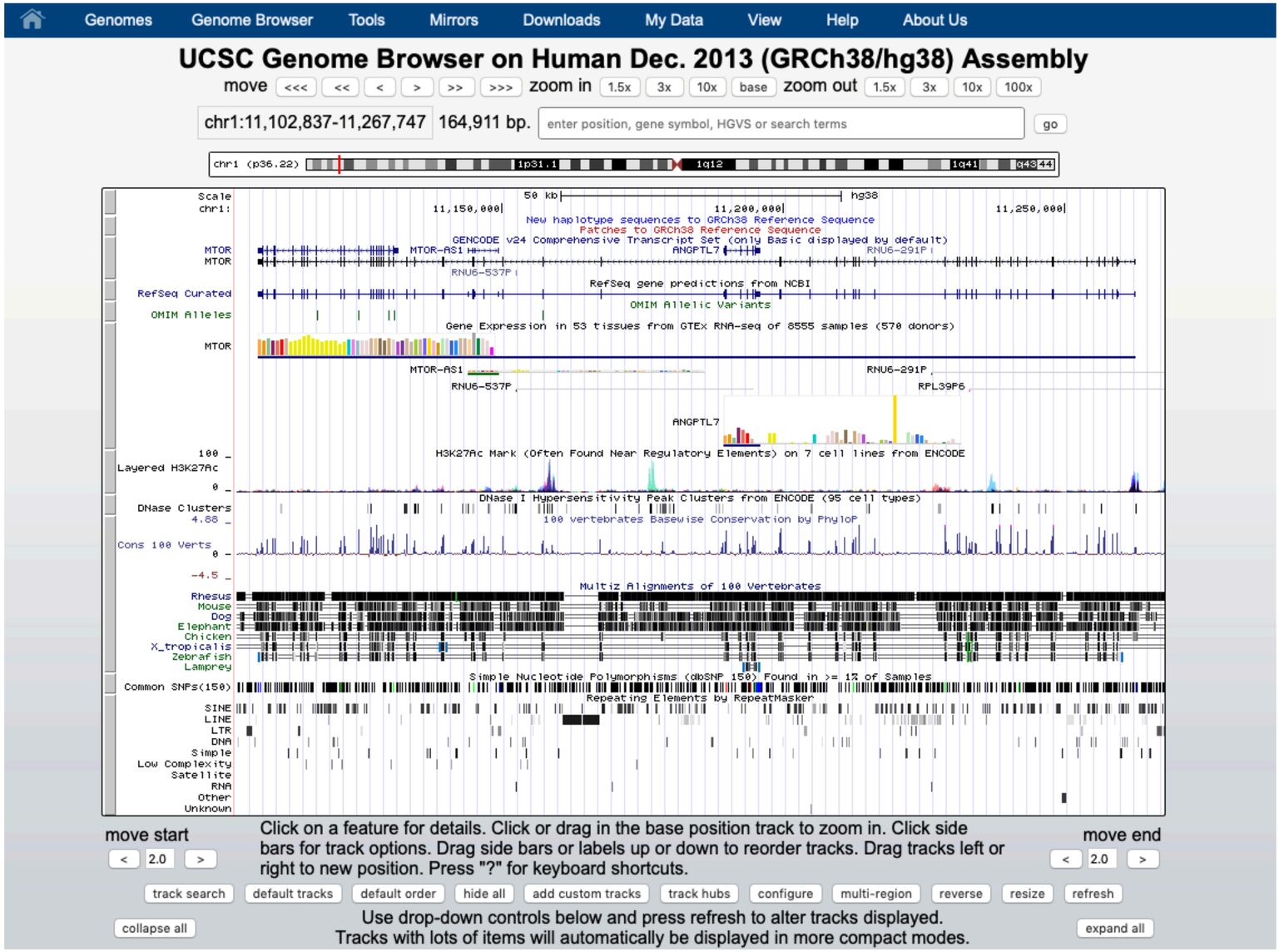
University of Zurich[UZH]

Reference Genome Resources...

RESOURCES FOR GENOMICS: UCSC GENOME BROWSER

- Originated from the Human Genome Project
- Most widely used general genome browser
- many default tracks
- many species
- customization with "BED" files

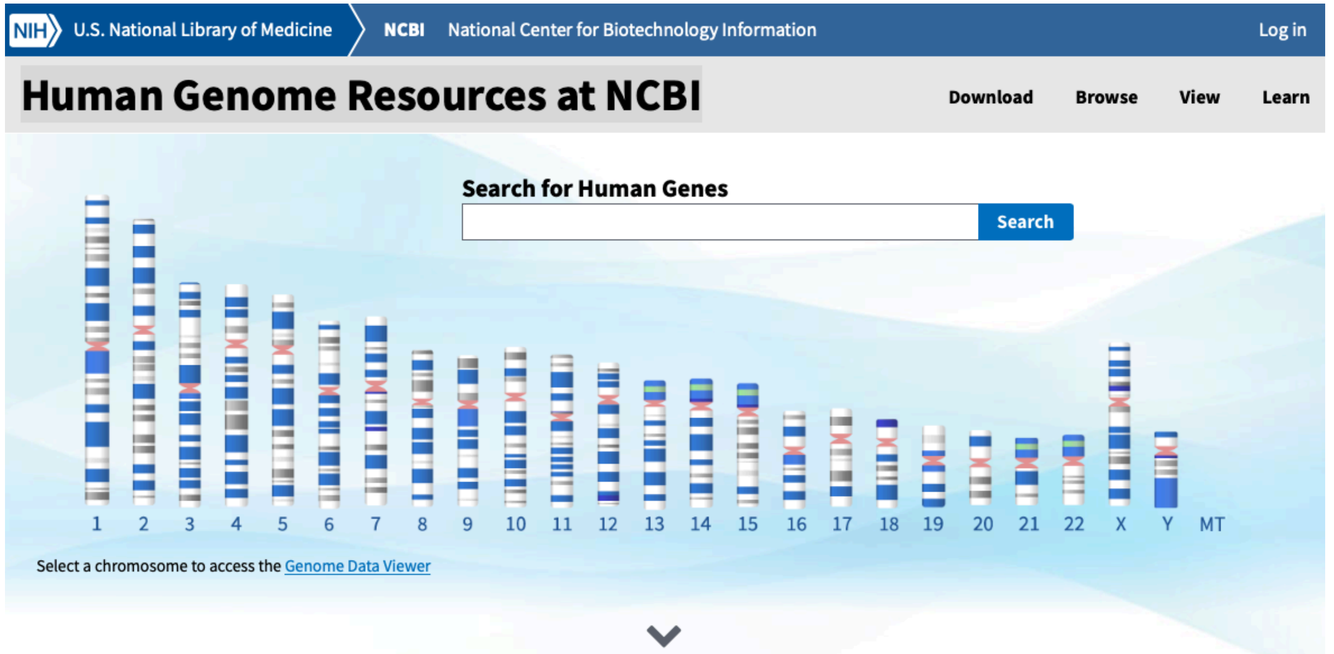
genome.ucsc.edu



RESOURCES FOR GENOMICS: HUMAN GENOME RESOURCES AT NCBI

- Entry point for genome reference data
- Human genome assemblies
- Human variant collections (dbVar, ClinVar, dbSNP) for download

www.ncbi.nlm.nih.gov/projects/genome/guide/human/

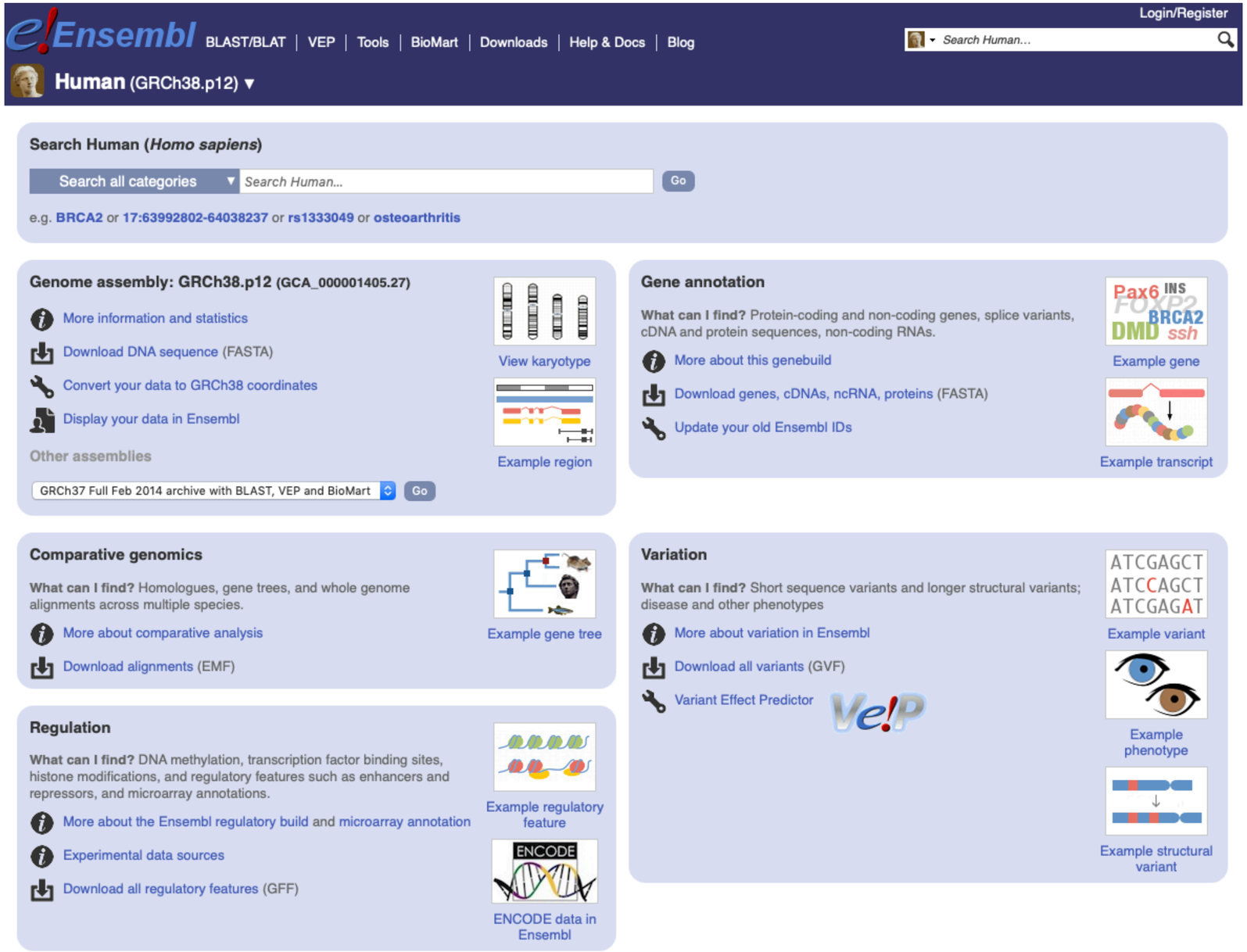


Download	GRCh38	GRCh37
Reference Genome Sequence	Fasta	Fasta
RefSeq Reference Genome Annotation	gff3	gff3
RefSeq Transcripts	Fasta	Fasta
RefSeq Proteins	Fasta	Fasta
ClinVar	vcf	vcf
dbSNP	vcf	vcf
dbVar	vcf	vcf

RESOURCES FOR GENOMICS: ENSEMBL

- Entry point for many genome data services and collections
- Downloads ("BioMart"), REST API

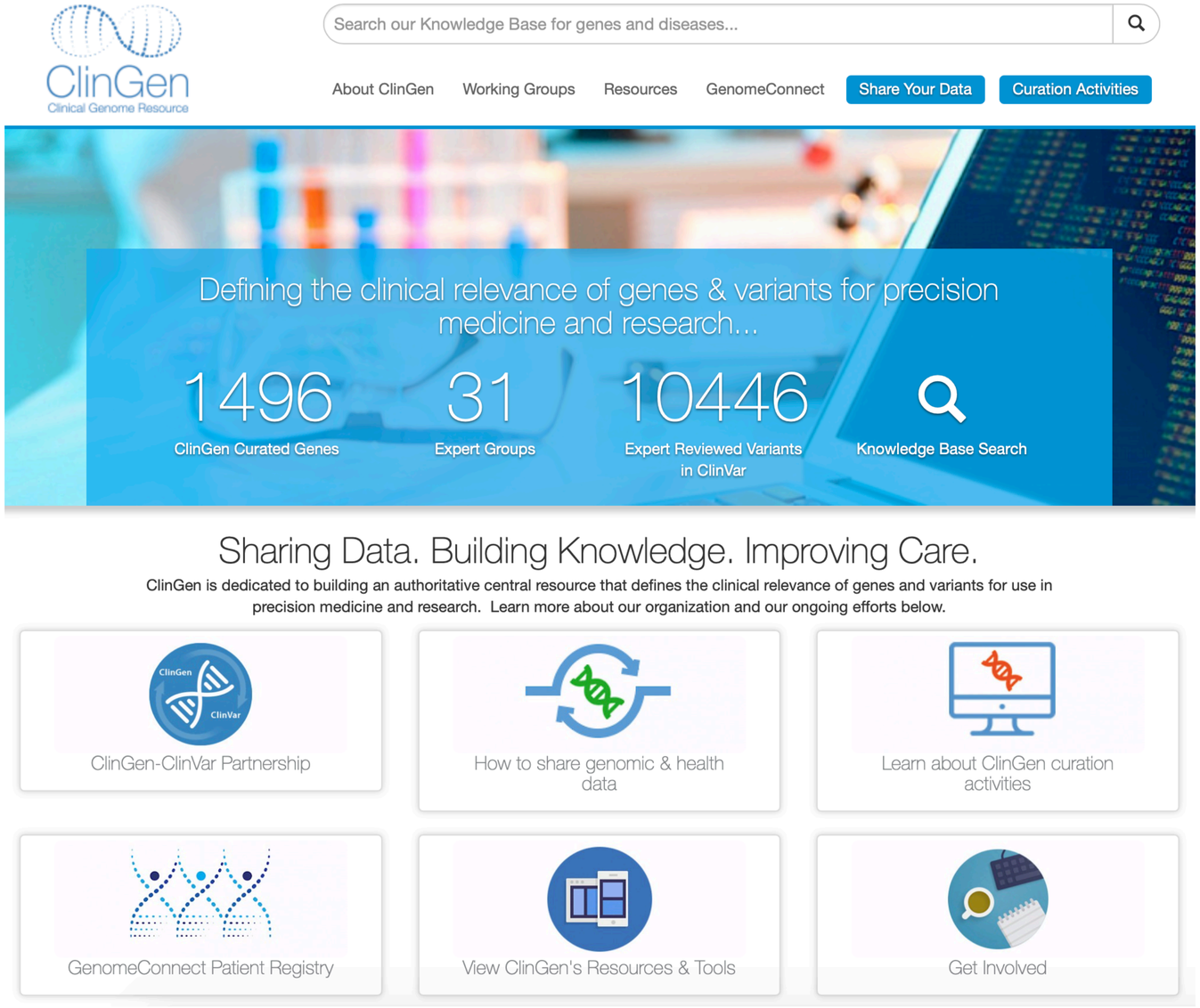
www.ensembl.org/Homo_sapiens/Info/Index



RESOURCES FOR GENOMICS: CLINGEN

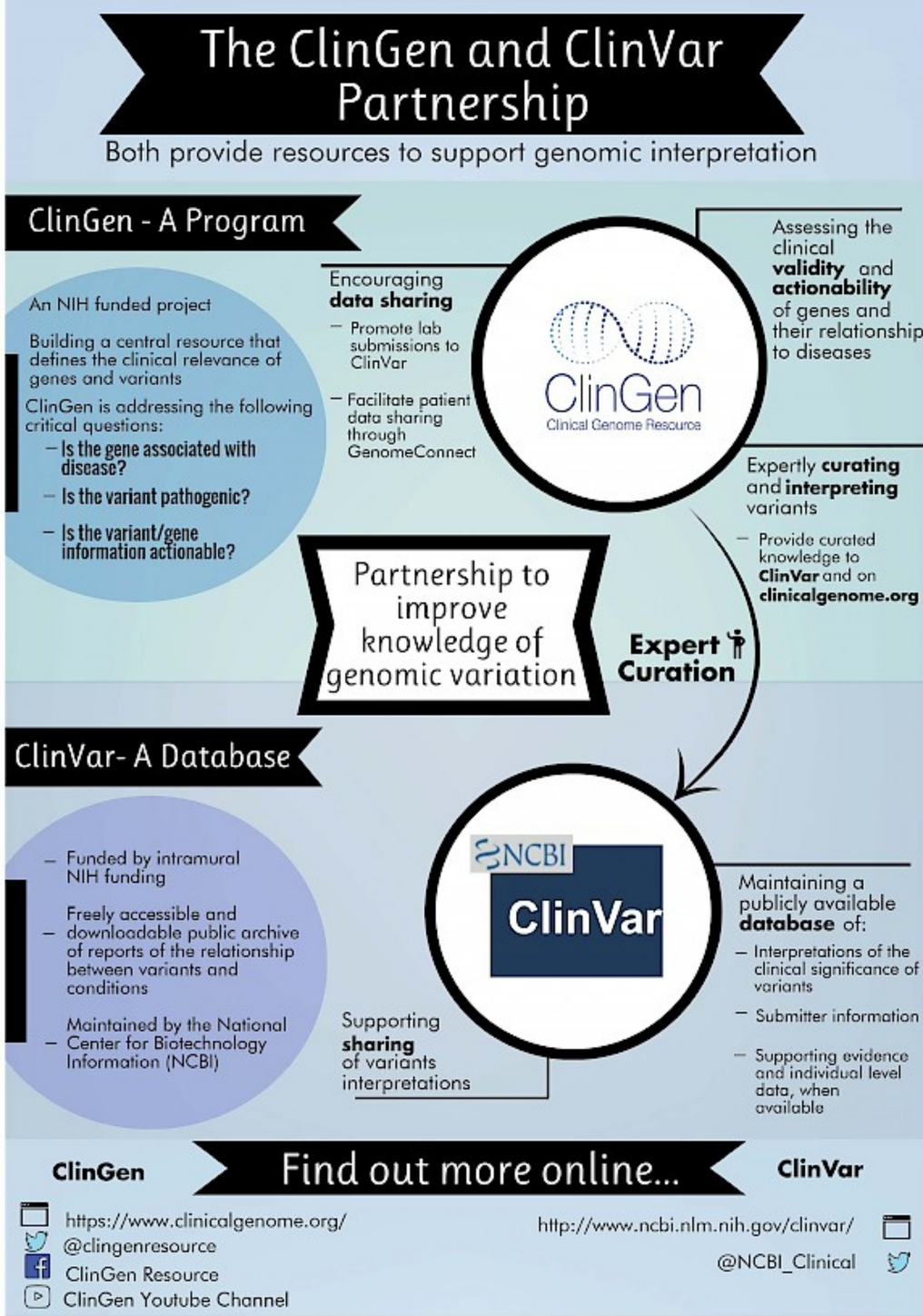
- "The Genomic Variant WG brings together representatives from the Sequence and Structural Variant communities for focused discussions on resolving discrepancies in variant interpretation and creating consistent curation guidelines."
- Interpreted genome variants with disease association

clinicalgenome.org



- ClinVar (an NCBI databbase/ resource) is used as basis for curated variant <-> disease associations in ClinGen
- ClinGen - a funded project (application/funding limited)
- ClinVar - an internal NIH resource (dependent on political "goodwill")

clinicalgenome.org



RESOURCES FOR CANCER GENOMICS

COSMIC
Catalogue of somatic mutations in cancer

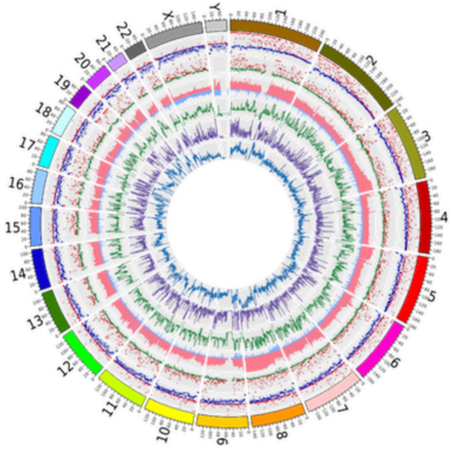
Home ▼ | Resources ▼ | Curation ▼ | Tools ▼ | Data ▼ | News ▼ | Help ▼ | About ▼ | Search COSMIC... | Login ▼

COSMIC v79, released 14-NOV-16

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below, or by browsing a region of the human genome using the map to the right.

eg: Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH



Browse the **genomic landscape** of cancer

R Resources

Key COSMIC resources

Cell Lines Project
COSMIC
Whole Genomes
Cancer Gene Census
Drug Sensitivity
Mutational Signatures
GRCh37 Cancer Archive

T Tools

Additional tools to explore COSMIC

Cancer Browser
Genome Browser
GA4GH Beacon
CONAN

C Expert Curation

High quality curation by expert postdoctoral scientists

Drug Resistance New
Cancer Gene Census
Curated Genes
Gene Fusions
Genome-Wide Screens

D Data

Further details on using COSMIC's content

Downloads
License
Submission
Genome Annotation
Datasheets
Help
FAQ

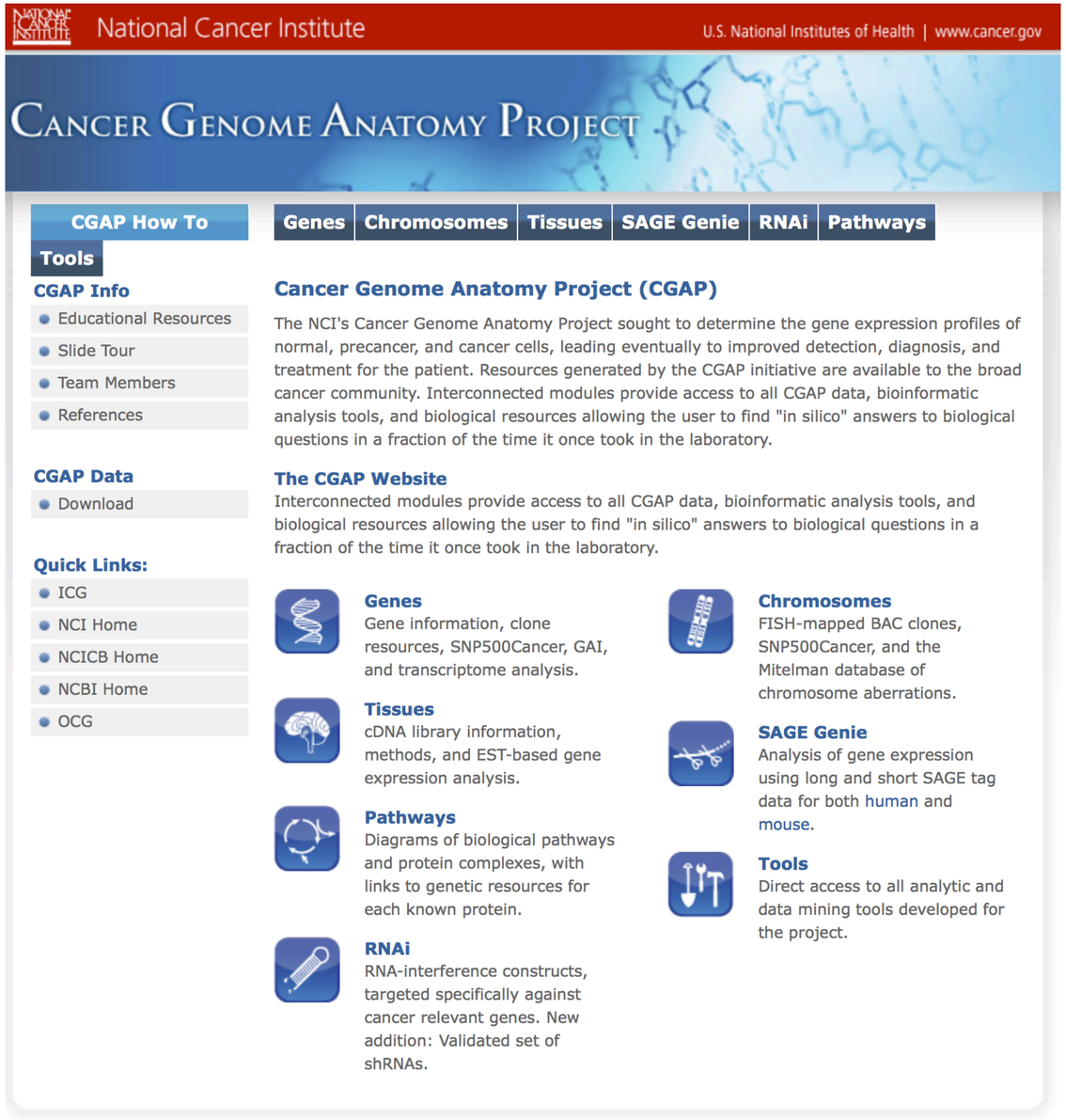
Cancer Gene Census Update

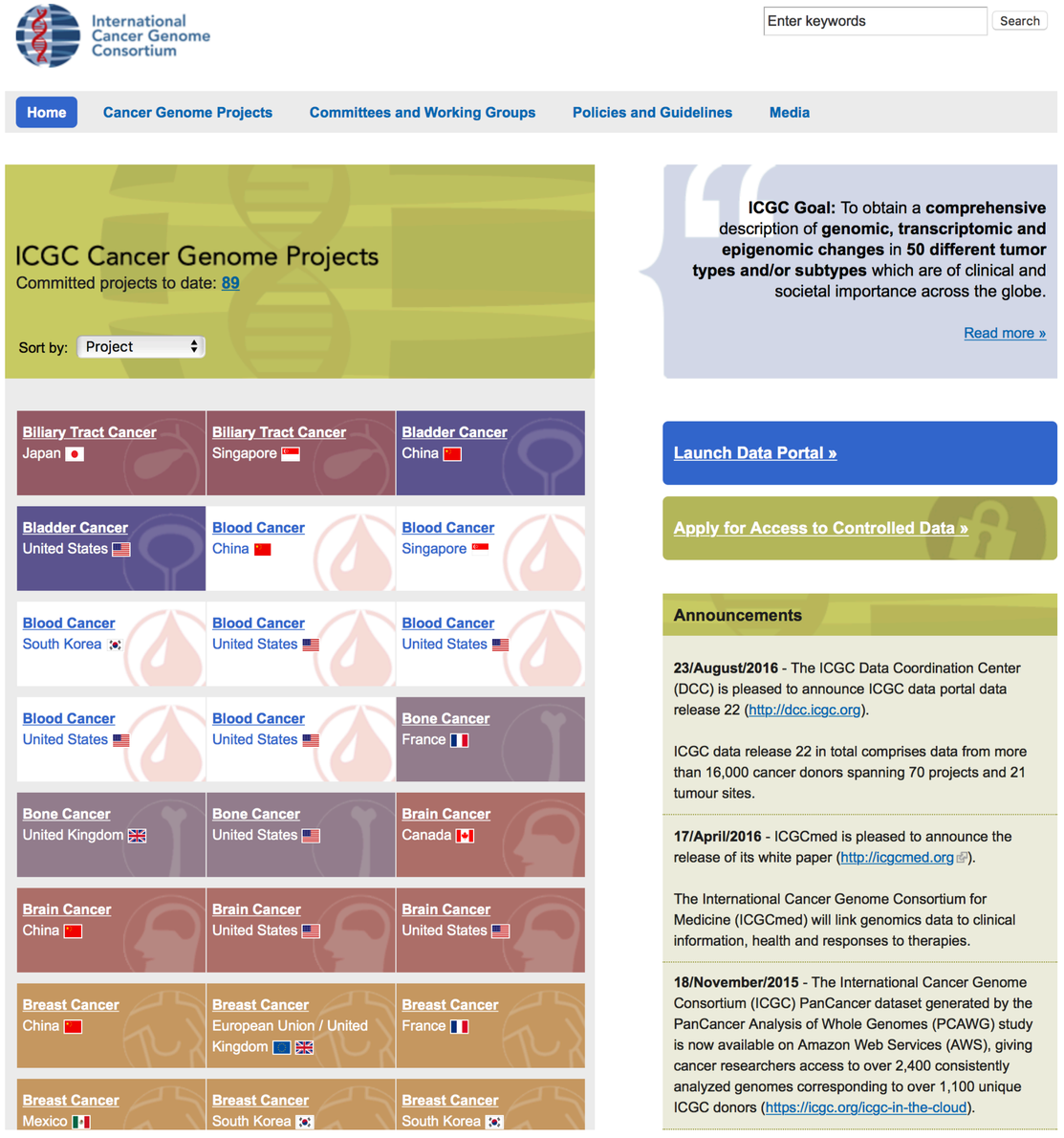
7 genes have been added to the Cancer Gene Census -

EPAS1 - Endothelial PAS domain protein 1.
PTPRT - Protein tyrosine phosphatase, receptor type T.
PPM1D - Protein phosphatase, Mg2+/Mn2+ dependent 1D.
BTK - Bruton tyrosine kinase.
PREX2 - Phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2.
TP63 - Tumour protein p63.
QKI - QKI, KH domain containing RNA binding.

For full details, see the Datasheet.

RESOURCES FOR CANCER GENOMICS



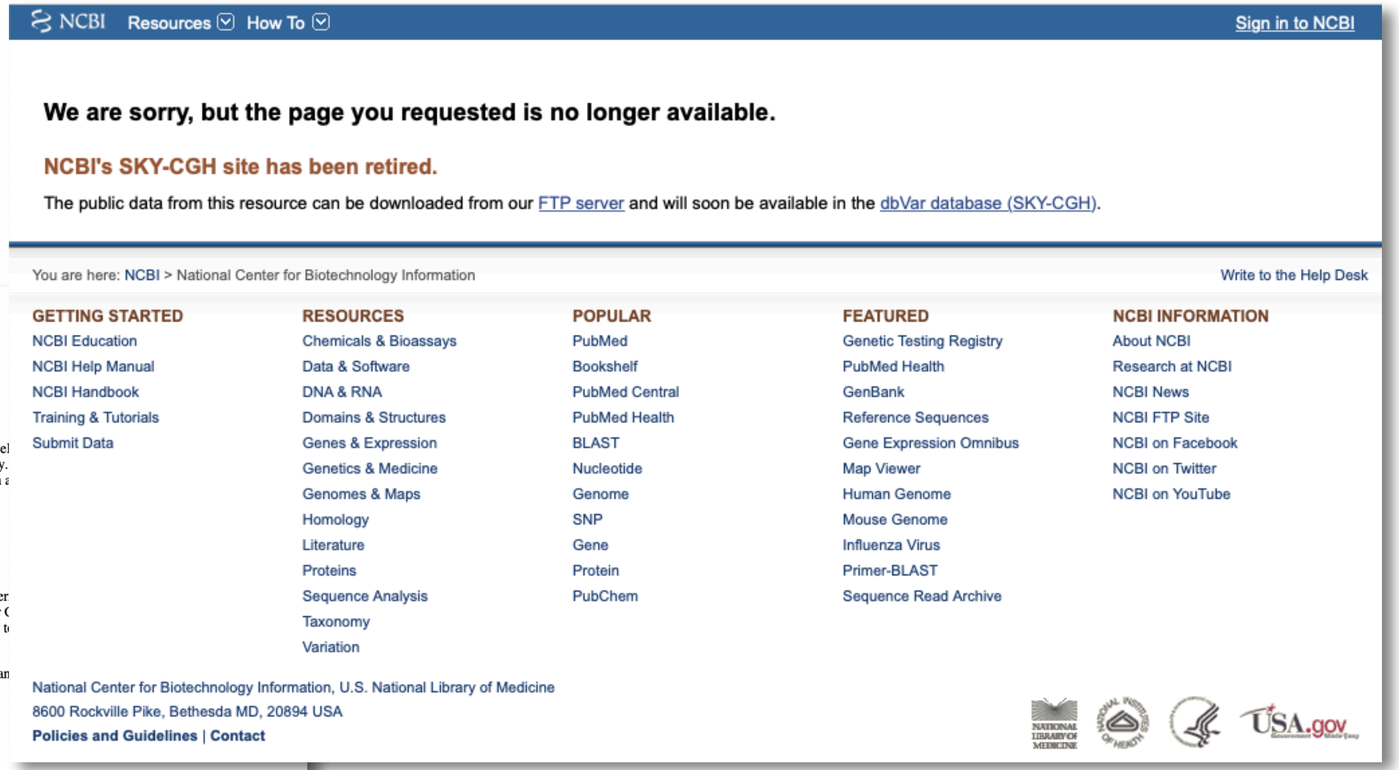


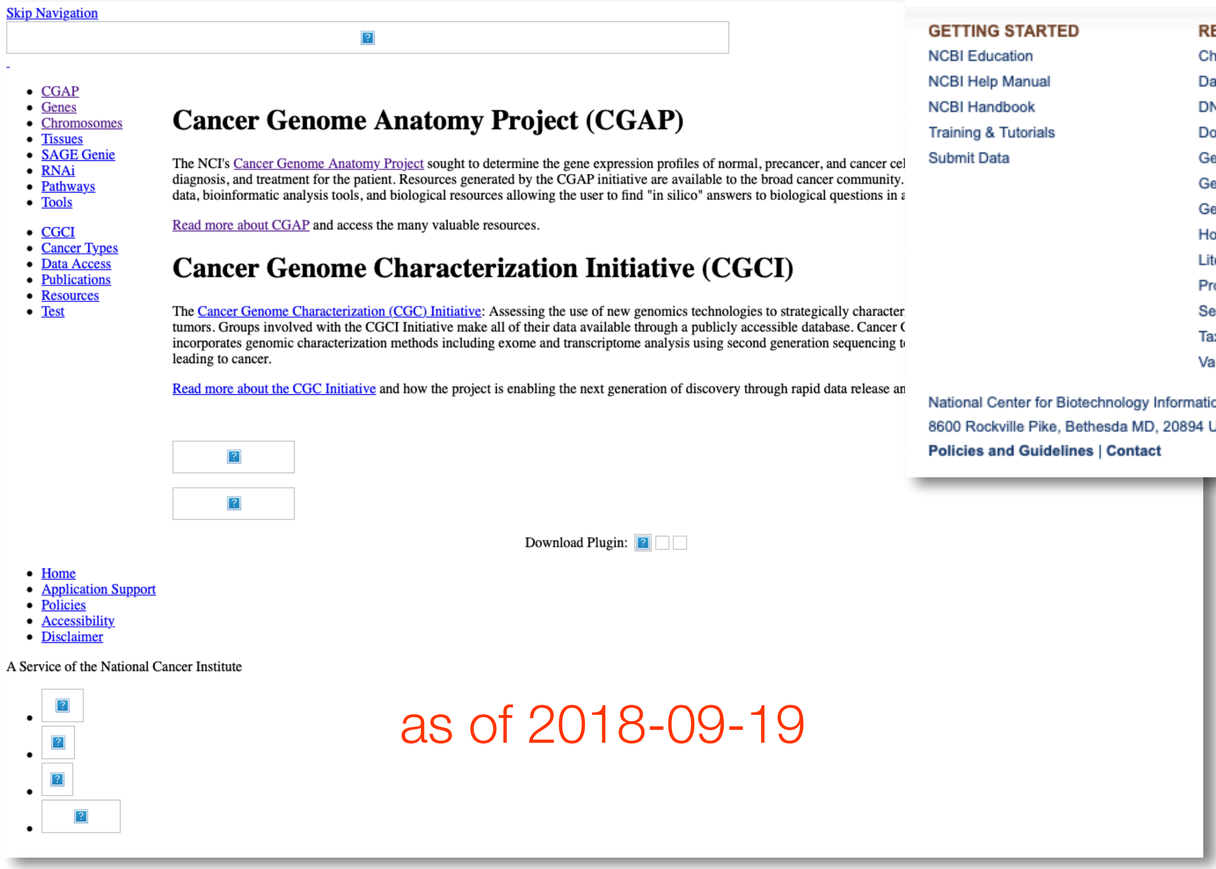
VARIANT RESOURCES FOR CANCER GENOMICS

Resource name	Primary institute	Constituent Knowledge base	Cancer focused	Therapeutic evidence	Predisp. evidence	Diagnostic evidence	Prognostic evidence	Variant emphasis	URL
Cancer Genome Interpreter (CGI)	Institute for Research in Biomedicine, Barcelona, Spain	x	x	x				Somatic	https://www.cancergenomeinterpreter.org/home
Clinical Interpretation of Variants in Cancer (CIViC)	Washington University School of Medicine (WashU)	x	x	x	x	x	x	All variants	www.civicdb.org
JAX Clinical Knowledgebase (CKB)	The Jackson Laboratory	x	x	x	x	x	x	Somatic	https://ckb.jax.org/
Molecular Match	Molecular Match	x	x	x			x	Somatic	https://app.molecularmatch.com/
OncoKB	Memorial Sloan Kettering Cancer Center	x	x	x				Somatic	http://oncokb.org/#/
Precision Medicine Knowledgebase (PMKB)	Weill Cornell Medical College	x	x	x	x	x	x	Somatic	https://pmkb.weill.cornell.edu/
BRCA exchange	GA4GH	x	x		x			Germline	http://brcaexchange.org/
Cancer Driver Log (CanDL)	Ohio State University (OSU) / James Cancer Hospital		x	x				Somatic	https://candl.osu.edu/
Gene Drug Knowledge Database	Synapse		x	x		x	x	Somatic	https://www.synapse.org/#!Synapse:syn2370773/wiki/62707
MatchMiner	Dana-Farber Cancer Institute		x					Somatic	http://bcb.dfci.harvard.edu/knowledge-systems/
COSMIC Drug Resistance Curation	Wellcome Trust Sanger Institute		x	x				Somatic	http://cancer.sanger.ac.uk/cosmic/drug_resistance
My Cancer Genome	Vanderbilt University		x	x		x	x	Somatic	https://www.mycancergenome.org/
NCI Clinical Trials	National Cancer Institute of the National Institutes of Health		x					Somatic	www.cancer.gov/about-cancer/treatment/clinical-trials
Personalized Cancer Therapy Database	MD Anderson Cancer Center		x	x	x	x	x	Somatic	https://pct.mdanderson.org/#/home
ClinGen Knowledge Base	ClinGen				x			Germline	https://www.clinicalgenome.org/resources-tools/
ClinVar	National Center for Biotechnology Information			x	x			All variants	http://www.ncbi.nlm.nih.gov/clinvar/
Pharmacogenomics Knowledgebase (PharmGKB)	Stanford University			x				Germline	https://www.pharmgkb.org/
The Human Gene Mutation Database (HGMD)	Institute of Medical Genetics in Cardiff				x			Germline	http://www.hgmd.cf.ac.uk

Wagner et al. A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. Nature Genetics (under review 2018)

RESOURCES FOR GENOMICS - THEY MAY BREAK SOMETIMES ...

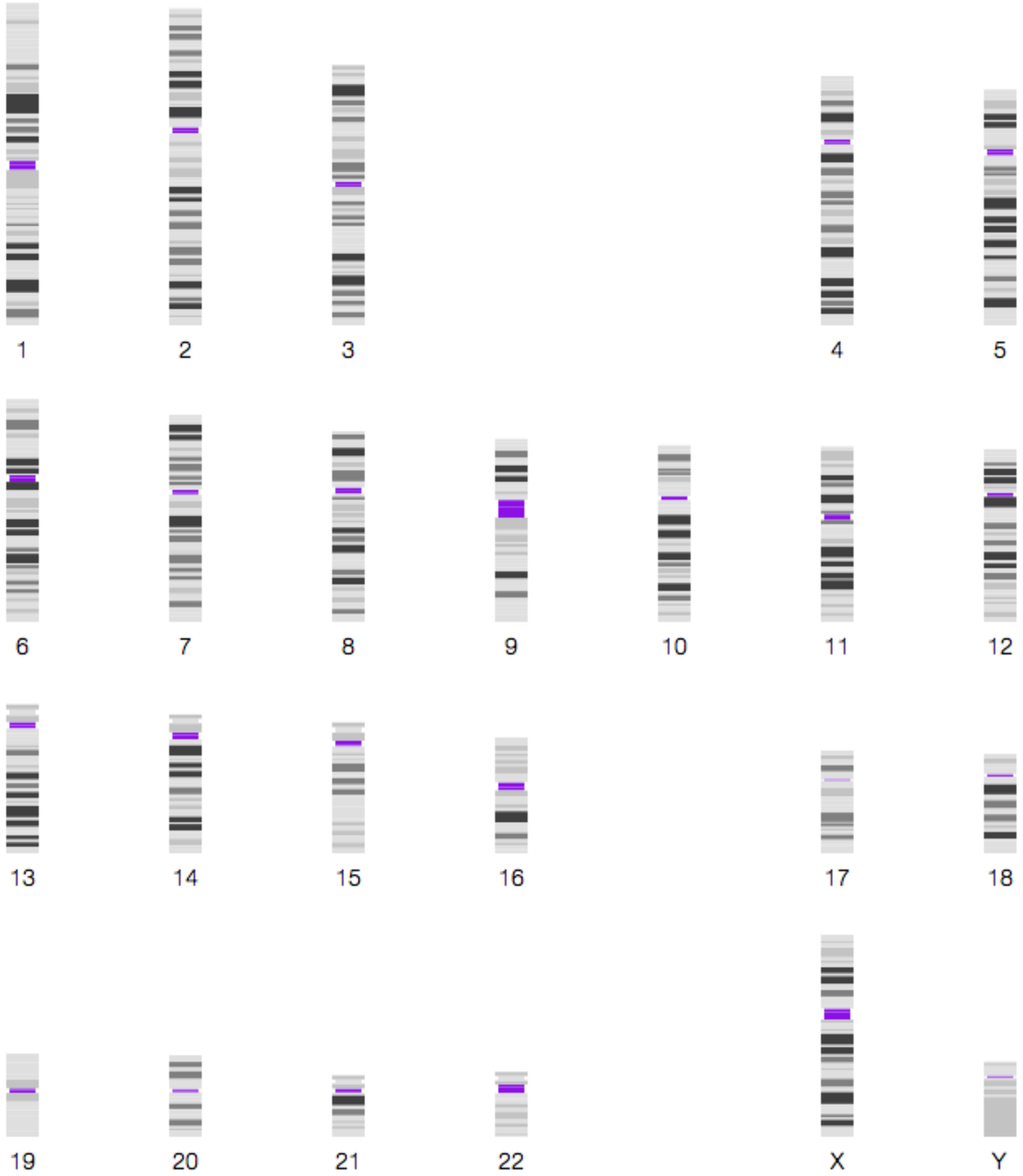




as of 2018-09-19

Genome Editions

Sizes | positions | mappings



Chromosome	Basepair length (GRCh38)
1	248'956'422
2	242'193'529
3	198'295'559
4	190'214'555
5	181'538'259
6	170'805'979
7	159'345'973
8	145'138'636
9	138'394'717
10	133'797'422
11	135'086'622
12	133'275'309
13	114'364'328
14	107'043'718
15	101'991'189
16	90'338'345
17	83'257'441
18	80'373'285
19	59'128'983
20	64'444'167
21	46'709'983
22	50'818'468
X	156'040'895
Y	57'227'415
	3'080'419'480

genome.ucsc.edu

cytoBand_UCSC_hg18.txt

chr1	0	2300000	p36.33	gneg
chr1	2300000	5300000	p36.32	gpos25
chr1	5300000	7100000	p36.31	gneg
chr1	7100000	9200000	p36.23	gpos25
chr1	9200000	12600000	p36.22	gneg
chr1	12600000	16100000	p36.21	gpos50
chr1	16100000	20300000	p36.13	gneg
chr1	20300000	23800000	p36.12	gpos25
chr1	23800000	27800000	p36.11	gneg
chr1	27800000	30000000	p35.3	gpos25
chr1	30000000	32200000	p35.2	gneg
chr1	32200000	34400000	p35.1	gpos25
chr1	34400000	39600000	p34.3	gneg
chr1	39600000	43900000	p34.2	gpos25
chr1	43900000	46500000	p34.1	gneg
chr1	46500000	51300000	p33	gpos75
chr1	51300000	56200000	p32.3	gneg
chr1	56200000	58700000	p32.2	gpos50
chr1	58700000	60900000	p32.1	gneg
...	...	...	...	...
chrX	130300000	133500000	q26.2	gpos25
chrX	133500000	137800000	q26.3	gneg
chrX	137800000	140100000	q27.1	gpos75
chrX	140100000	141900000	q27.2	gneg
chrX	141900000	146900000	q27.3	gpos100
chrX	146900000	154913754	q28	gneg
chrY	0	1700000	p11.32	gneg
chrY	1700000	3300000	p11.31	gpos50
chrY	3300000	11200000	p11.2	gneg
chrY	11200000	11300000	p11.1	acen
chrY	11300000	12500000	q11.1	acen
chrY	12500000	14300000	q11.21	gneg
chrY	14300000	19000000	q11.221	gpos50
chrY	19000000	21300000	q11.222	gneg
chrY	21300000	25400000	q11.223	gpos50
chrY	25400000	27200000	q11.23	gneg
chrY	27200000	57772954	q12	gvar

Cytogenetic band sizes

chromosome	band start position	band stop position	cytogenetic band	staining intensity	band size
chr6	63400000	63500000	q11.2	gneg	100000
chr15	64900000	65000000	q22.32	gpos25	100000
chr17	22100000	22200000	p11.1	acen	100000
chrX	65000000	65100000	q11.2	gneg	100000
chrY	11200000	11300000	p11.1	acen	100000
chr17	35400000	35600000	q21.1	gneg	200000
chr3	44400000	44700000	p21.32	gpos50	300000
chr3	51400000	51700000	p21.2	gpos25	300000
chr9	132500000	132800000	q34.12	gpos25	300000
chr13	45900000	46200000	q14.13	gneg	300000
chr15	65000000	65300000	q22.33	gneg	300000
chr1	120700000	121100000	p11.2	gneg	400000
chr8	39500000	39900000	p11.22	gpos25	400000
chr9	72700000	73100000	q21.12	gneg	400000
chr16	69400000	69800000	q22.2	gpos50	400000
chr19	43000000	43400000	q13.13	gneg	400000
chr9	70000000	70500000	q13	gneg	500000
chr20	41100000	41600000	q13.11	gneg	500000
...	...	...	...	...	...
chr9	51800000	60300000	q11	acen	8500000
chrX	76000000	84500000	q21.1	gpos100	8500000
chr11	76700000	85300000	q14.1	gpos100	8600000
chr13	77800000	86500000	q31.1	gpos100	8700000
chr7	77400000	86200000	q21.11	gpos100	8800000
chr8	29700000	38500000	p12	gpos75	8800000
chr3	14700000	23800000	p24.3	gpos100	9100000
chr5	82800000	91900000	q14.3	gpos100	9100000
chr6	104800000	113900000	q21	gneg	9100000
chrX	120700000	129800000	q25	gpos100	9100000
chr9	60300000	70000000	q12	gvar	9700000
chr1	212100000	222100000	q41	gpos100	10000000
chr1	128000000	142400000	q12	gvar	14400000
chr1	69500000	84700000	p31.1	gpos100	15200000
chrY	27200000	57772954	q12	gvar	30572954

Positional genomic data has to be evaluated in the context of the correct edition

Chromosome	Basepairs 2003 (HG16)	Basepairs 2006 (HG18)	Basepairs 2009 (HG19)	Basepairs 2013 (GRCh38)	HG16 => HG19
1	246'127'941	247'249'719	249'250'621	248'956'422	2'828'481
2	243'615'958	242'951'149	243'199'373	242'193'529	-1'422'429
3	199'344'050	199'501'827	198'022'430	198'295'559	-1'048'491
4	191'731'959	191'273'063	191'154'276	190'214'555	-1'517'404
5	181'034'922	180'857'866	180'915'260	181'538'259	503'337
6	170'914'576	170'899'992	171'115'067	170'805'979	-108'597
7	158'545'518	158'821'424	159'138'663	159'345'973	800'455
8	146'308'819	146'274'826	146'364'022	145'138'636	-1'170'183
9	136'372'045	140'273'252	141'213'431	138'394'717	2'022'672
10	135'037'215	135'374'737	135'534'747	133'797'422	-1'239'793
11	134'482'954	134'452'384	135'006'516	135'086'622	603'668
12	132'078'379	132'349'534	133'851'895	133'275'309	1'196'930
13	113'042'980	114'142'980	115'169'878	114'364'328	1'321'348
14	105'311'216	106'368'585	107'349'540	107'043'718	1'732'502
15	100'256'656	100'338'915	102'531'392	101'991'189	1'734'533
16	90'041'932	88'827'254	90'354'753	90'338'345	296'413
17	81'860'266	78'774'742	81'195'210	83'257'441	1'397'175
18	76'115'139	76'117'153	78'077'248	80'373'285	4'258'146
19	63'811'651	63'811'651	59'128'983	59'128'983	-4'682'668
20	63'741'868	62'435'964	63'025'520	64'444'167	702'299
21	46'976'097	46'944'323	48'129'895	46'709'983	-266'114
22	49'396'972	49'691'432	51'304'566	50'818'468	1'421'496
X	153'692'391	154'913'754	155'270'560	156'040'895	2'348'504
Y	50'286'555	57'772'954	59'373'566	57'227'415	6'940'860
	3'070'128'059	**3'080'419'480**	**3'095'677'412**	**3'088'781'199**	**18'653'140**

Variant Annotation Formats

- Linear annotation of single-letter **nucleotides** or amino acid codes
- leading information line, usually with unique SeqID
- text format
  - "readable"
  - not optimised for size
- representation of a sequence without ambiguities or QC data
  - extended as "FASTQ" (Sanger Centre)

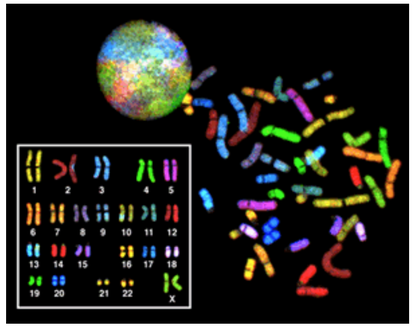
```
>NC_000007.14:11369935-11832697 Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly
AGGGCTTAAATGGTTCCCTACTTACTTACATTTTAGCAAATAGCTATTTCAGAAAATGTTTTAACTGCAA
ACTACCCCGGAAGTAACTTGTCTTAAGTTGTGTGTCCCTTCCTGAATTTGTTAAGGCATAAGTTTCTGCT
TTGACTTTTAGGTTTGGTTTTTTGTGGGTAGACACAGGGACAAGAGACAGTGAGGGATGTGCCTTATGAC
TGATTGGGTGGGAAAAGCTGTACTCTGTTGTAGAGAGTTTCCCCATCCTCTGCTGCTGCCATTTTTGAAT
TGACTGGAAACCAGGAGGTCCCTTGTCCATGATTCACCTGGTGGCCTCAGCCAACTTTCAAAGTAAAAGT
TTGCATTTCTTGAACTTTCTAAATTGGAGTTGTTATACAACCCAGGAAAGGGCAATACAGTAGGTAAAAG
GATTTAGGTATTCACTGGAAAAAAATTTTAATCCATATTTAAGAAGCAATTGGTGCAAATCAAAACACAG
ATACACATGATTAGAATGAAAATGATTTCCGTATTTATGTTTGTCAGCAATATAGTTATTTACAAATAAC
CCATATGAAAATGTAAAAAAGCATATTACATCTTCACATGCCATCTGTATTGACTGAATAAAGCTTAGTG
ACATTATTTGCAAATCTGTAGTTAATTGTACATAGACATTTTGTGCGTTAAAAAGGAAATGTACATAATG
TAAAATAAATTACATTACGCAATTTACAAAGTAATATTAACAAAAATTCTTAGACAGCTGCCTCCTTATT
TAAACAAAATAAATATTCAGGTAGTTAAATTAACATAAAATACACATTTAGGGAATAATAATTTATAGAA
AGACAGATTGCAAATTTTAATAAAGTTATATTTTACAATGATAGATACTGATCTCTTCAAATCTGTGTGA
TAGAAATGGAGAAAAAAAGTACCAGAAAAGGAATCTAAATGTTTACTTTCTAAAAATAAACACAAACAGA
TTTCTGAAAAATAGGGAAAAGTTTACTTGAGGGTTAAAAGTAGGTAATCTAGAAACTATGGCTTAAAAAC
AATAAATCTACAAAACACAACAAGACTGACAATTATATTCTAAATAATATAGAGATTTTTGATCACTGAA
AACATGACTCCCACAAACTAAAGCTCTTCATATACTGCCCATTTTTAAGATCTGACTTGGTAGAAACACA
GAAAAATAAAAATGCAAAATTAACTGTTTAGCATTAGTTTCATTCATTTAATAATGTAGACATAACCATT
TTTCATTGTCCCTGCTAGATATAAAATTATTATAACACACTGCAAACCATCCTTTTTATAAATGGATAAC
TATTTGCTTGGCTCACACACCAGTTTCTGATACCTGAAATCCTGTCCTGCAGCCAGGGCCACCTGAGGGC
AGGACCTGGGGAGACCCTTATTCCCAGAACAAGCAGATGTAGTTCTCACAACTAAACTAGTCCCAGGAAA
GATCACATTCTGACAAGATTCCTCACAGATTTGCTCAAGGACTACTGTTTTTTCAACACCCTCAATCTTA
CAGTGGAAATAGAAGAAGAACCCACACTTGAATTTGGTAATATATTTATAAACAGGAGATCCCAGATCAT
TTGGGAATTGTGCTTCTCATGTACTATTGAGACCCACGTCAGCTTTAGAACAGGCTCTCCCTTCTGTATG
GTACTGAAAAGTACAGTCCTCCCTCACTGTCTTGTGGTGTGGAATGTGAACAAAGACTCGAGATGGAGGC
AGGAGGATATGGGATGGTCTAAAGCAAGTGTAGGCATGGACATTTTTACAGAAAAGGGCTTTTTTTTTTT
TTTTTTTCCTGCATGCCTTCCACATTTTTCCTTTTCCCTTTATTCATTTCTTTGACCAGTGGATTTGGTG
...
```

Homo sapiens chromosome 7, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000007.14

GENOMIC VARIANT FORMATS: ISCN

- ISCN - "International System for Human Cytogenetic Nomenclature"
- Annotation format for chromosomal aberrations, i.e. traditional microscopically visible structural and quantitative abnormalities in karyotypes
- extensions for "molecular cytogenetics" (e.g. M-FISH, SKY, genomic arrays)



SKY - Spectral Karyotyping of tumour metaphase (source: https://www.genome.gov)

Symbol	Description
,	Separates modal number (total number of chromosomes), sex chromosomes, and chromosome abnormalities
-	Loss of a chromosome
( )	Grouping for breakpoints and structurally altered chromosomes
+	Gain of a chromosome
;	Separates rearranged chromosomes and breakpoints involving more than one chromosome
/	Separates cell lines or clones
//	Separates recipient and donor cell lines in bone marrow transplants
del	Deletion
der	Derivative chromosome
dic	Dicentric chromosome
dn	*de novo* (not inherited) chromosomal abnormality
dup	Duplication of a portion of a chromosome
fra	Fragile site (usually used with Fragile X syndrome)
h	Heterochromatic region of chromosome
i	Isochromosome
ins	Insertion
inv	Inversion
.ish	Precedes karyotype results from FISH analysis
mar	Marker chromosome
mat	Maternally-derived chromosome rearrangement
p	Short arm of a chromosome
pat	Paternally-derived chromosome rearrangement
psu dic	*pseudo dicentric* - only one centromere in a Dicentric chromosome is active
q	Long arm of a chromosome
r	Ring chromosome
t	Translocation
ter	Terminal end of arm (e.g. 2qter refers to the end of the long arm of chromosome 2)
tri	Trisomy
trp	Triplication of a portion of a chromosome

Source: Wikipedia (https://en.wikipedia.org/wiki/Cytogenetic_notation)

GENOMIC VARIANT FORMATS: DBVAR

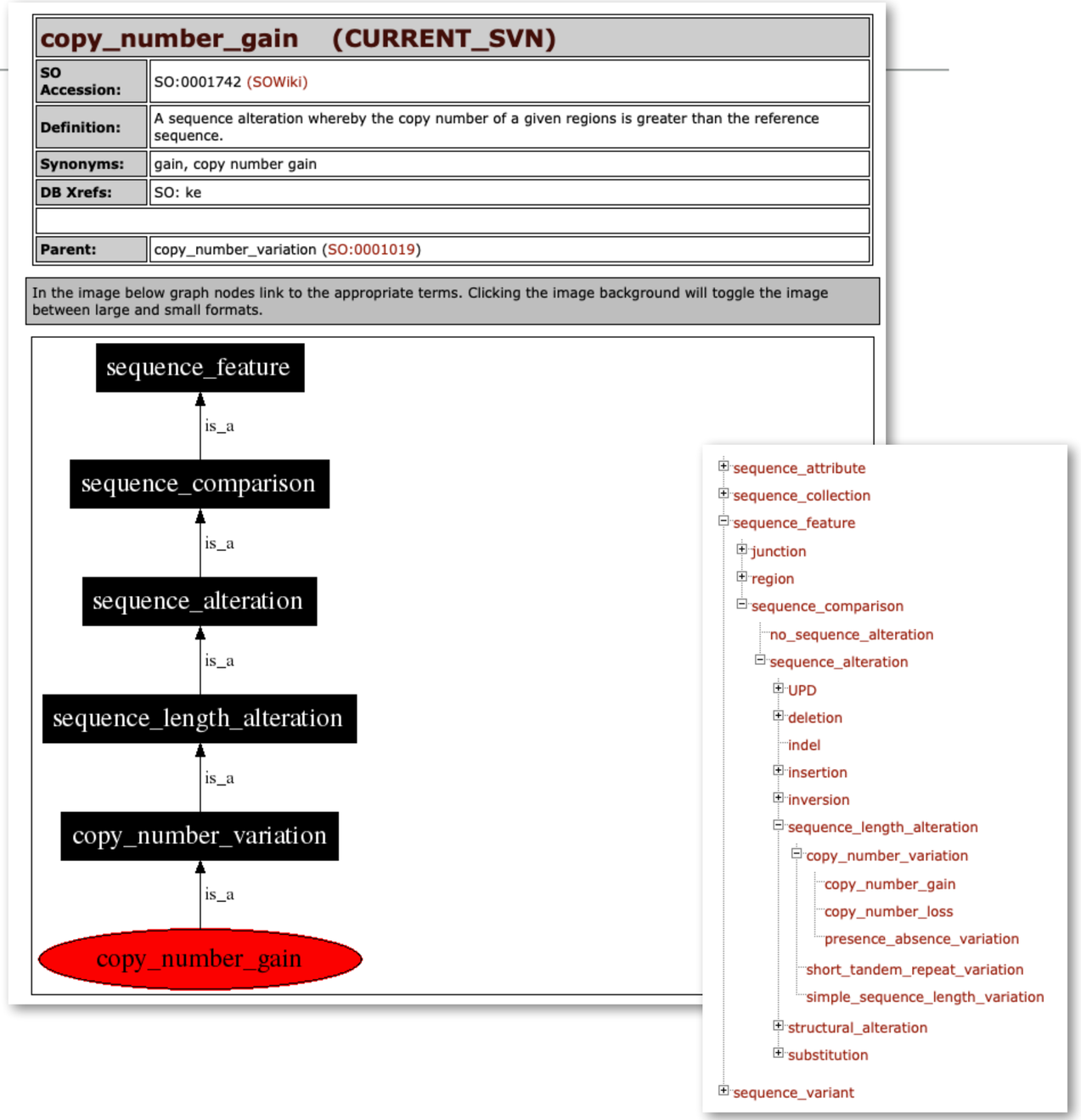
- dbVar is "NCBI's database of human genomic structural variation — insertions, deletions, duplications, inversions, mobile elements, and translocations"
- structural genome variations are still not completely solved with respect to unambiguous annotation

ncbi.nlm.nih.gov/dbvar/content/overview/

Variant Call Type	Sequence Ontology ID	Variant Region Type
copy number gain	**SO:0001742** A sequence alteration whereby the copy number of a given region is greater than the reference sequence.	copy number variation
copy number loss	**SO:0001743** A sequence alteration whereby the copy number of a given region is less than the reference sequence.	copy number variation
duplication	**SO:0001742** (copy number gain) A sequence alteration whereby the copy number of a given region is greater than the reference sequence.	copy number variation
deletion	**SO:0000159** The point at which one or more contiguous nucleotides were excised.	copy number variation
insertion	**SO:0000667** The sequence of one or more nucleotides added between two adjacent nucleotides in the sequence.	insertion
mobile element insertion	**SO:0001837** A kind of insertion where the inserted sequence is a mobile element.	mobile element insertion
novel sequence insertion	**SO:0001838** An insertion the sequence of which cannot be mapped to the reference genome.	novel sequence insertion
tandem duplication	**SO:1000173** A duplication consisting of 2 identical adjacent regions.	tandem duplication
inversion	**SO:1000036** A continuous nucleotide sequence is inverted in the same position.	inversion
intrachromosomal breakpoint	**SO:0001874** A rearrangement breakpoint within the same chromosome.	translocation or complex chromosomal mutation
interchromosomal breakpoint	**SO:0001873** A rearrangement breakpoint between two different chromosomes.	translocation or complex chromosomal mutation
translocation	**SO:0000199** A region of nucleotide sequence that has translocated to a new position.	translocation
complex	**SO:0001784** A structural sequence alteration or rearrangement encompassing one or more genome fragments.	complex
sequence alteration	**SO:0001059** A sequence_alteration is a sequence_feature whose extent is the deviation from another sequence.	sequence alteration
short tandem repeat variation	**SO:0002096** A kind of sequence variant whereby a tandem repeat is expanded or contracted with regard to a reference.	short tandem repeat variation

Source: Wikipedia (https://en.wikipedia.org/wiki/Cytogenetic_notation)

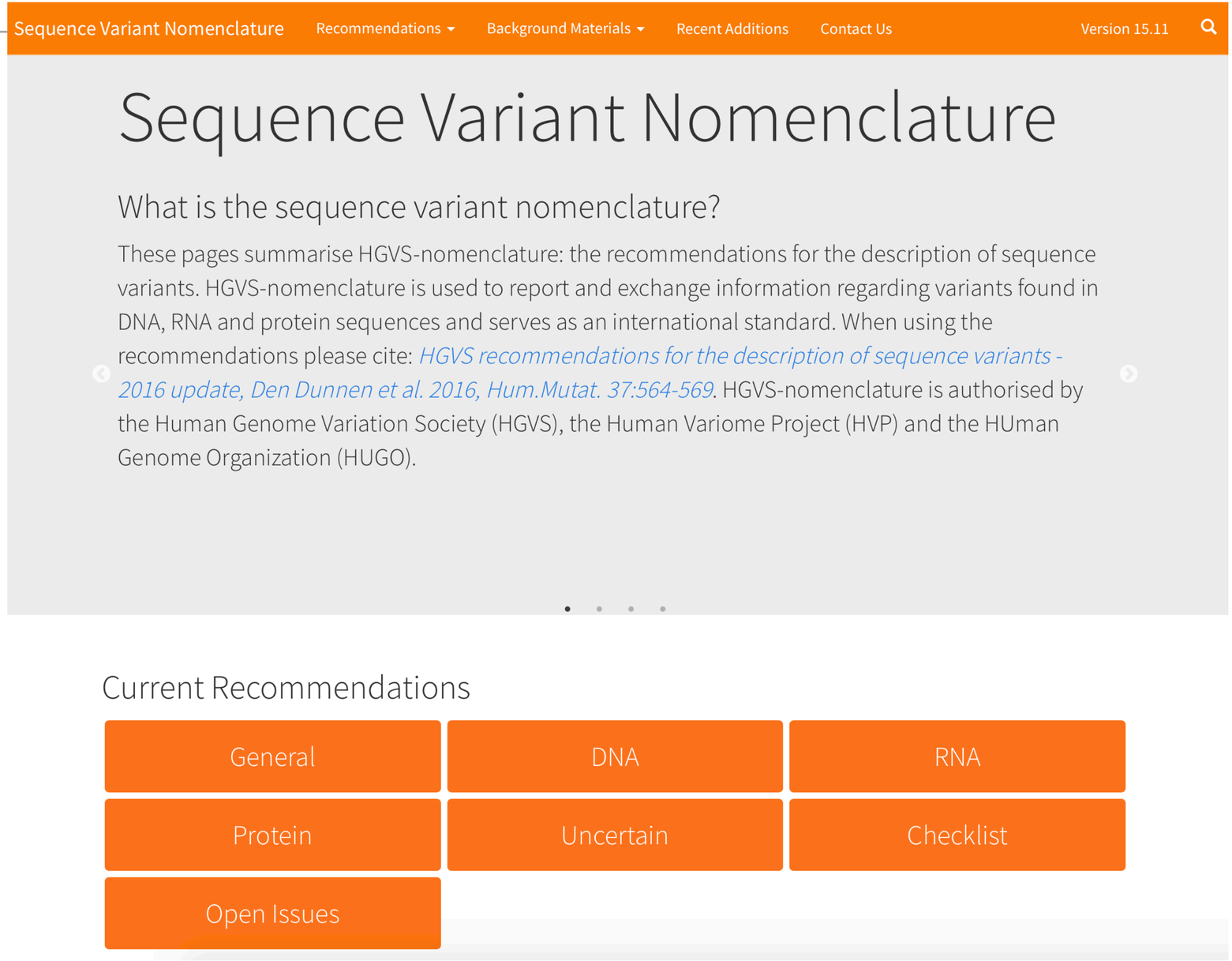
- Sequence Ontology describes types of biological sequence alterations (or normal status)
- It is by itself not suitable for complete variant description (e.g. lacking the localisation; has to be attached to a sequence or functional element)



The Sequence Ontology

GENOMIC VARIANT FORMATS: HGVS

- "The consistent and unambiguous description of sequence variants is essential to report and exchange information on the analysis of a genome. In particular, DNA diagnostics critically depends on accurate and standardized description and sharing of the variants detected. The sequence variant nomenclature system proposed in 2000 by the Human Genome Variation Society has been widely adopted and has developed into an internationally accepted standard."



- HGVS allows the annotation of sequence variants (DNA, RNA, protein) with relation to a genomic ("g") or protein ("c") reference

HGVS Variation Examples

A Single Nucleotide Variant : rs268

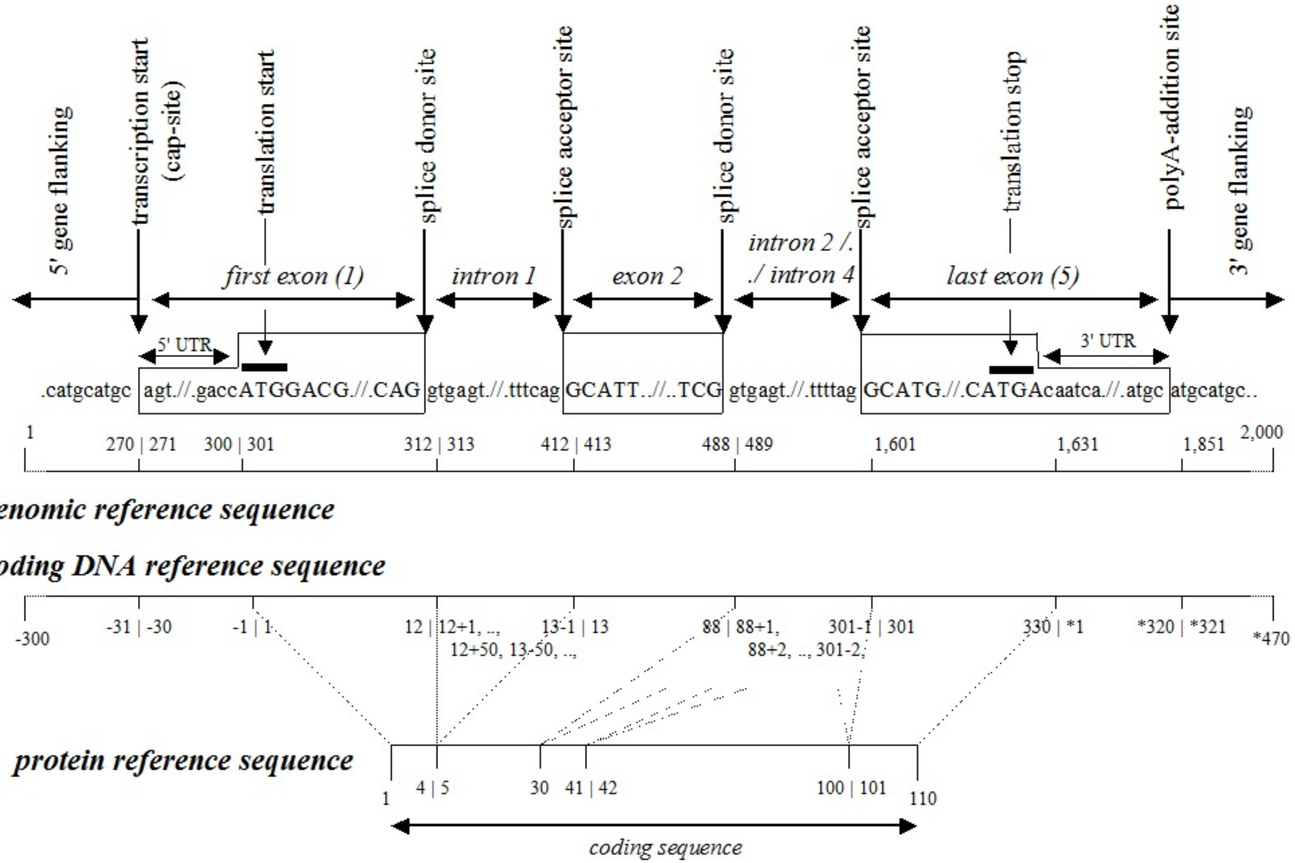
- NC_000008.10:g.19813529A>G
- NG_008855.1:g.21948A>G
- NM_000237.2:c.953A>G
- NP_000228.1:p.Asn318Ser

An Insertion Variant : rs9281300

- NC_000006.11:g.31239170_31239171insA
- NG_029422.2:g.5738_5739insT
- NM_001243042.1:c.344-46_344-45insT
- NM_002117.5:c.344-46_344-45insT

A Deletion Variant : rs1799758

- NC_000016.9:g.2138200_2138203delTGAG
- NG_005895.1:g.43894_43897delTGAG
- NM_000548.3:c.5161-28_5161-25del



Source NCBI:
https://www.ncbi.nlm.nih.gov/variation/tools/reporter/docs/examples

http://www.hgvs.org/mutnomen/refseq_figure.html

All of these are the same variant. Or not.

NC_000001.10:g.103471457_103471459delCAT (ClinVar Id 93966)
= NC_000001.10:g.103471486_103471488delTCA
Right shifted per HGVS Nomenclature guidelines

NM_001166478.1:c.30_31insT
= NM_001166478.1:c.35dupT
Normalized and rewritten

NM_080588.2:c.139C>G (rs4073458)
= ENST00000367279:c.139C>G
Has identical CDS and exon structure, including UTR

NP_003768.2:p.(Arg4412Alafs*2) (rs72658833)
= NP_003768.2:p.(Arg4412Alafs)
= NP_003768.2:p.(Arg4412AlaTrpTer)
Same protein truncation (+ wo/parens and 1-letter forms!)

CC BY

Reece Hart, 2016

"The simplest thing that might work."

```
"vmc:allele": {
    "reference_sequence_id": "NCBI:NM000059.3",
    "interval": {"start": 50, "end": 51},
    "edit": "A"
},

"vmc:genotype": {
    "alleles": [
        {
            "reference_sequence_id": "NCBI:NM000059.3",
            "interval": {"start": 50, "end": 51},
            "edit": "A",
        },
        {
            "reference_sequence_id": "NCBI:NM000059.3",
            "interval": {"start": 50, "end": 51},
            "edit": "T",
        }
]},
```

CC BY

Or…


```
{
   "vmc:alleles": [{
      “id”: “VA_5e632de6e7280769”,
      "reference_sequence_id": "VS_451ec666acc937f1",
      "interval": {"start": 50, "end": 51},
      "alternate": "A"
      }, (more alleles)
        ],

   "vmc:genotypes": [{
      “id”: “VG_5e632de6e7280769”,
      "allele_ids": [“VA_5e632de6e7280769”, “VA_72802de6e7695e63”]
      }, (more genotypes)
      ]

   "vmc:haplotypes": [{
      “id”: “VH_de8d7b851fb84223”,
      "allele_ids": [“VA_5e632de6e7280769”, “VA_d7b851fb84223de8”]
      }, (more haplotypes)
   ]

   "vmc:diplotype": [{
      “id”: “VD6fd159c94192f252”,
      "haplotype_ids": [“VH_de8d7b851fb84223”, “VH_b851fb8de8d74223”],
      }, (more diplotypes)
   ]
}
```


CC BY

VARIANT ANNOTATION FORMAT: THE "GA4GH TRANSITIONAL" MODEL

- object model for the representation of genomic variants in the context of an experimental read-out ("callset")
- based on VCF principles (e.g. notation for structural variants such as "DUP", DEL" ...)
- allows to place variants into intervals with "fuzzy ends"
- not yet suitable for genotype reconstruction (e.g. connecting translocations, other out-of-place events)

https://github.com/ga4gh-metadata/schemas

79 lines (78 sloc) | 3.32 KB — Raw | Blame | History

```yaml
info:
  title: |
    GA4GH __variant__
  description: |
    The document describes attributes of the _variant_ object. In its current implementation, _valiant_ (and related genomic o
  updated: '2018-09-14'
definitions:
  Variant:
    properties:
      digest:
        description: concatenated unique specific elements of the variant
        type: string
        example: '4:12282-46465:DEL'
      callset_id:
        description: The identifier ("callset.id") of the callset this variant is part of.
        type: string
        example: 'PGX_AM_CS_GSM1690424'
      biosample_id:
        description: The identifier ("biosample.id") of the biosample this variant was reported from. This is a shortcut to us
        example: 'pgx-bs-987647'
      reference_name:
        description: 'Reference name (chromosome). Accepting values 1-22, X, Y.'
        type: string
        example: "8"
      start:
        description: array of 1 or 2 (for imprecise end position of structural variant) integers
        type: array
        format: int64
        example: [20867740]
      mate_name:
        description: 'Mate name (chromosome) for fusion (BRK) events; otherwise left empty. Accepting values 1-22, X, Y.'
        type: string
        example: "14"
      end:
        description: array of 0 (for presise sequence variants), 1 or 2 (for imprecise end position of structural variant) int
        type: array
        format: int64
        example: [21977798,21978106]
        queries:
          - query: 'db.variants.find( { "reference_name" : "9",  "variant_type" : "DEL", "start" : { $lteq : 21975098 }, "end"
            description: the query will return all variants with any overlap of the CDKN2A CDR
      reference_bases:
        description: one or more bases at start position in the reference genome, which have been replaced by the alternate_ba
        type: string
        pattern: '^([ACGT]+|N)$'
        example: 'G'
      alternate_bases:
        description: one or more bases relative to start position of the reference genome,replacing the reference_bases value;
        type: string
        pattern: '^([ACGT]+|N)$'
        example: 'AC'
      genotype:
        description: list of strings, which represent the (phased) alleles in which the variant was being observed
        type: array
        example:
          - '1'
          - '.'
      variant_type:
        description: the variant type in case of a named (structural) variant (e.g. DUP, DEL, BRK ...)
        type: string
        example: 'DEL'
      info:
        description: additional variant information, as defined in the example and accompanying documentation
        schema:
          $ref: './common/Info_class'
        example:
          cnv_value:
            type: number
            format: float
            value: -0.294
          cnv_length:
            type: number
            format: int64
            value: 1205290
      updated:
        description: time of the last edit of this record, in ISO8601
        type: string
        example: '2017-10-25T07:06:03Z'
```

University of Zurich[UZH]

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB I Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis

Global Alliance
for Genomics & Health